# The Bias-Variance Tradeoff: A Deep Dive

Nipun Batra and teaching staff

IIT Gandhinagar

August 15, 2025

# Table of Contents

# Understanding the Problem Setup

# The Learning Problem: A Real-World Example

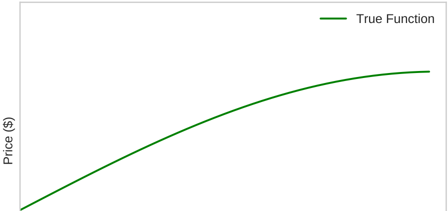
**Definition: Our Scenario**

**Goal:** Predict housing prices based on house area

**Example: The True Relationship**

**Unknown to us:** There exists a true function $f_{\theta_{\text{true}}}$ that perfectly relates area to price:

$$y_t = f_{\theta_{\text{true}}}(x_t)$$



# The Three Sources of Prediction Error

**Important: Fundamental Question**

**Why do our predictions fail?** What causes the difference between our predictions and reality?

**Definition: Three Universal Sources of Error**

**Every machine learning prediction suffers from:**

1. **Noise** - Irreducible randomness in the data
2. **Bias** - Systematic errors from model assumptions
3. **Variance** - Sensitivity to particular training sets

**Key Points**

**The Tradeoff:** We can often reduce bias OR variance, but not both simultaneously!

# Source 1: Noise - The Irreducible Error

# Understanding Noise: The Fundamental Limitation

**Definition: What is Noise?**

**Noise** represents factors affecting the target that we cannot observe or control

**Example: Real-World Noise Sources**

**In housing prices:**

- House condition (hard to measure precisely)
- Neighborhood market dynamics
- Buyer's personal preferences
- Economic conditions on sale day
- Unmeasurable aesthetic factors

**Important: Key Insight**

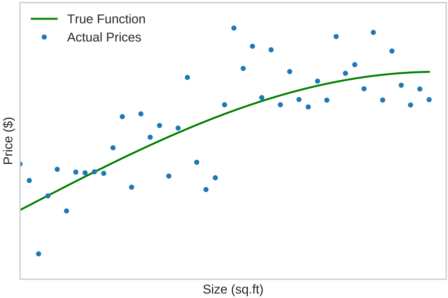
# Noise: Mathematical Formulation

**Key Points**

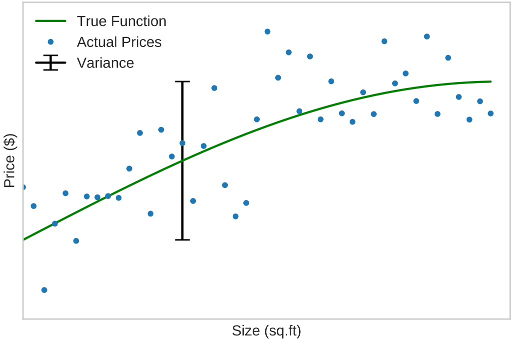
The Noisy Relationship **True relationship becomes:**

$$y_t = f_{\theta_{\text{true}}}(x_t) + \epsilon_t$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is the noise term



# Visualizing Noise: Data Distribution




**Key Points**

**Key Observation:**

- Data points scatter around the true function
- The spread (variance) is constant: $\sigma^2$

# Source 2: Bias - Systematic Model Limitations

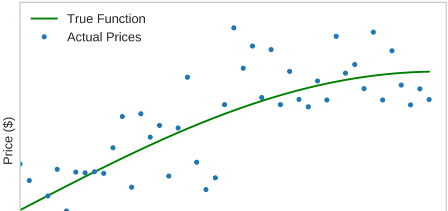
# Understanding Bias: Model Flexibility

**Definition: What is Bias?**

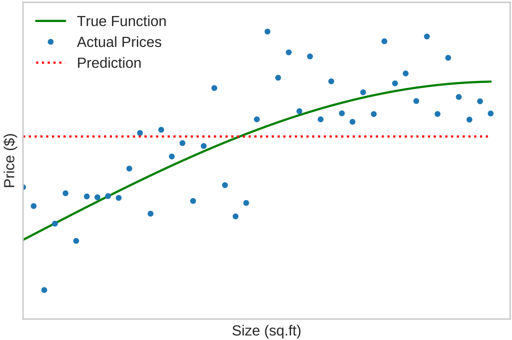
**Bias** measures how well our model class can represent the true function

**Example: Extreme Example: Constant Function**

**Model choice:** $\hat{f}(x) = c$ (constant, regardless of house size)
**Question:** Can this model capture the true price-size relationship?



# Bias: Fitting a Constant Model

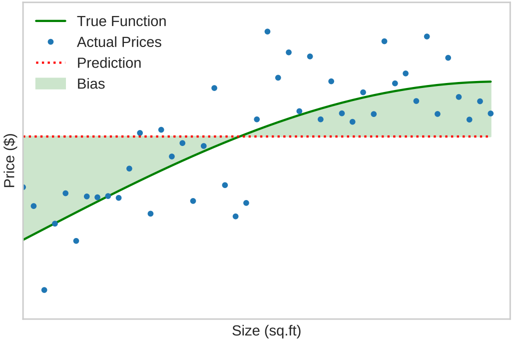



## Key Points

**Best Constant Fit:**

- The optimal constant is the average of all prices
- But this completely ignores the size information!

# Bias: Visualizing the Systematic Error




**Definition: Bias Definition**

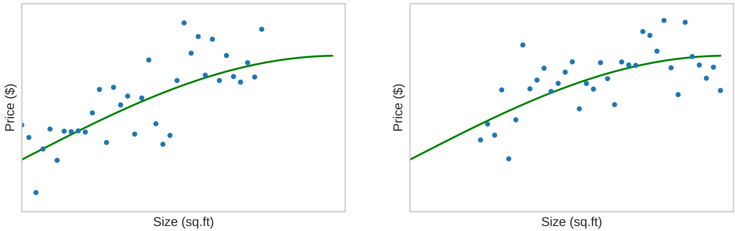
$$\text{Bias}(x) = f_{\theta_{\text{true}}}(x) - E[\hat{f}(x)]$$

**The systematic difference between truth and average prediction**

# Multiple Datasets: Understanding Variability

**Key Points**

**Crucial Insight:** Many different datasets are possible from the same true relationship!

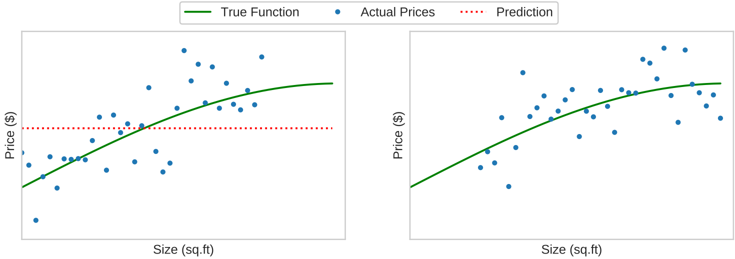



**Example:**

Same underlying relationship, different data points due to:

- Random sampling of houses

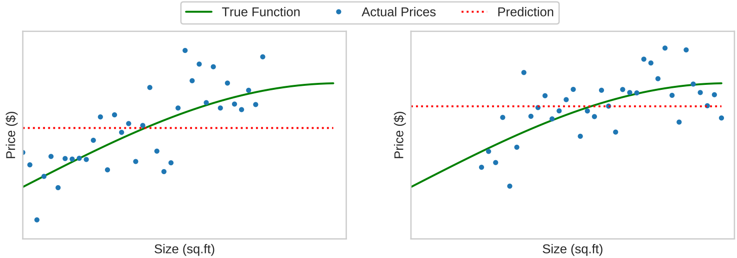
# Fitting Models to Different Datasets




**Key Points**

**Question:** If we fit the same model type (constant) to different datasets, what happens?

# Different Predictions from Different Datasets




**Important:**

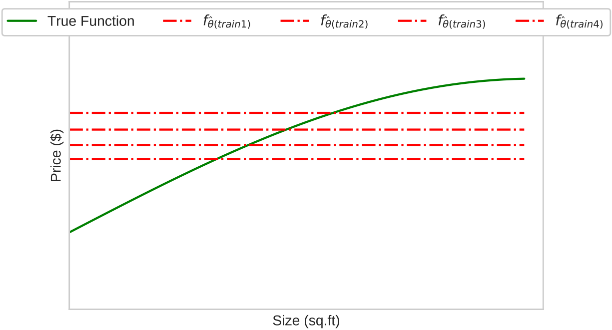
Key Observation: Even with the same model type, we get different predictions!

**Definition:**

This variability leads us to two concepts:

- **Average prediction:** What happens "on average" across all

# Many Datasets: The Full Picture



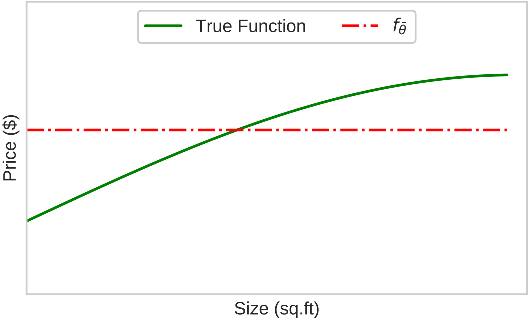

**Key Points**

**Multiple Datasets:** Each gives a slightly different constant fit

**Example:**

The Big Question: What is the "typical" or "expected" prediction

# The Average Model: Expected Prediction




**Definition: Expected Prediction**

$$E[\hat{f}(x)] = \text{Average prediction across all possible training sets}$$

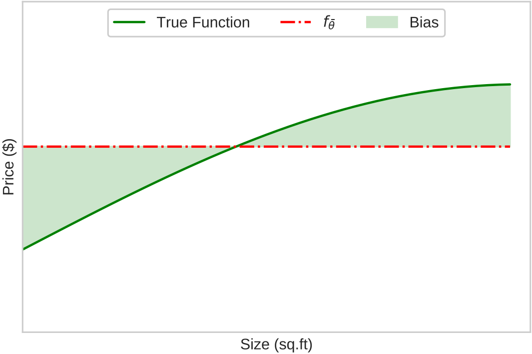
**Key Points**

# Bias: The Final Definition

**Definition: Bias Formula**

$$\text{Bias}(x) = f_{\theta_{\text{true}}}(x) - E[\hat{f}(x)]$$

**Difference between truth and expected prediction**

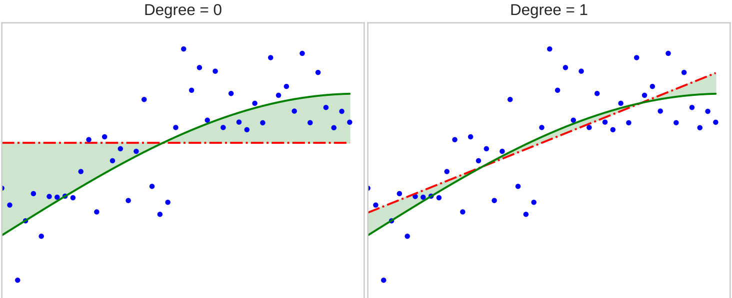


# Model Complexity vs Bias: The Relationship

## Key Points

**Universal Pattern:**

- **Increase complexity** $\rightarrow$ Model becomes more flexible
- **More flexible** $\rightarrow$ Can better approximate true function
- **Better approximation** $\rightarrow$ Bias decreases



# High-Complexity Models: Near-Zero Bias


**Important:**

High-Degree Polynomials: Can approximate almost any smooth function!

**Definition:**

# Source 3: Variance - Dataset Sensitivity

# From Bias to Variance: The Other Side

**Important:**

We've seen: High-complexity models have low bias
**Question:** If low bias is good, why not always use high-complexity models?

**Definition: Enter Variance**

**Variance** measures how much predictions change when we train on different datasets

**Key Points**

**Intuition:**

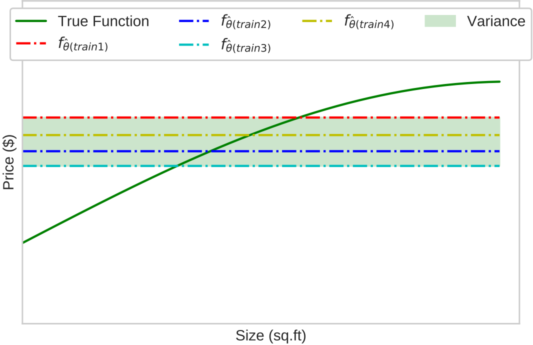
- Simple models: Stable, consistent predictions
- Complex models: Highly sensitive to specific training data

# Understanding Variance: Prediction Consistency

**Definition: Variance Definition**

**Variance** = How much do predictions vary across different training sets?

$$\text{Var}(\hat{f}(x)) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

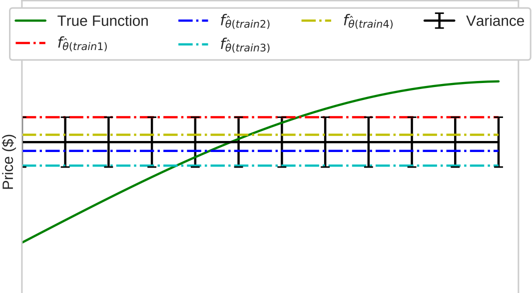


# Low Complexity: Low Variance

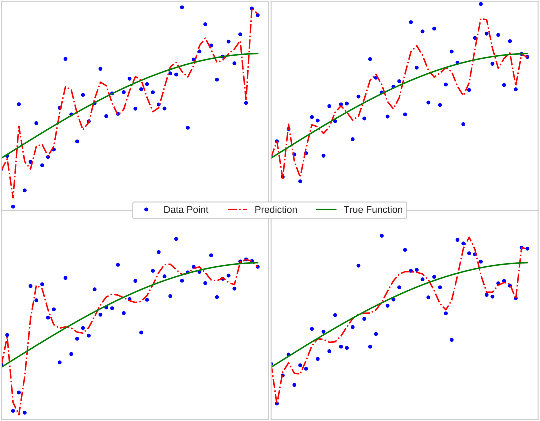
## Key Points

**Simple Models (e.g., linear):**

- Few parameters to estimate
- Robust to data variations
- Consistent predictions



# High Complexity: The Variance Problem Emerges




**Important:**

Warning Signs:

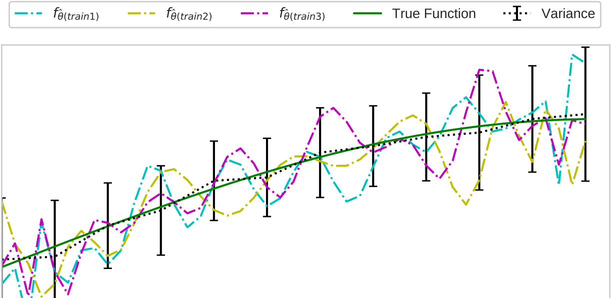
- Models look very different across datasets

# High Complexity: Extreme Variance

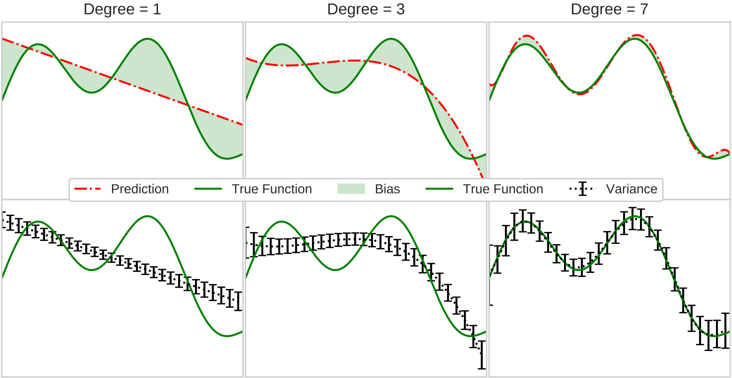
## Key Points

**Complex Models (e.g., high-degree polynomials):**

- Many parameters to estimate
- Overfit to specific training data
- Dramatically different predictions



# The Bias-Variance Tradeoff: The Central Tension




**Important: The Fundamental Tradeoff**

- **Simple models:** High bias, low variance
- **Complex models:** Low bias, high variance
- **Optimal complexity:** Balance between the two

# Mathematical Derivation: The Bias-Variance Decomposition

# Why Mathematical Analysis Matters

## Definition: The Goal

**Question:** Can we mathematically prove that error $=$ bias$^2$ $+$ variance $+$ noise?

## Key Points

Why This Matters

- **Theoretical foundation:** Understand the fundamental nature of learning
- **Model selection:** Know exactly what we're trading off
- **Algorithm design:** Create methods that explicitly balance bias and variance

## Example: Expected Error Across All Possible Datasets

# Setting Up the Mathematical Framework

**Definition: What We Want to Prove**

$$E[\text{Error}] = \text{Noise} + \text{Bias}^2 + \text{Variance}$$

**Key Points**

Our Approach

1. Start with prediction error at a single point
2. Use squared loss: $(y - \hat{f}(x))^2$
3. Take expectation over all sources of randomness
4. Apply algebraic manipulation to separate terms

**Example: Sources of Randomness**

- **Training set:** Which data points we observe

# Mathematical Setup: Defining the Components

**Definition: True Relationship with Noise**

$$y = f_{\text{true}}(x) + \epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

**Key Points**

Key Definitions

- $f_{\text{true}}(x)$: The unknown true function
- $\hat{f}(x)$: Our model's prediction (depends on training data)
- $E[\hat{f}(x)]$: Expected prediction over all possible training sets
- $\epsilon$: Irreducible noise with variance $\sigma^2$

# Formal Definitions: Bias and Variance

**Definition: Bias**

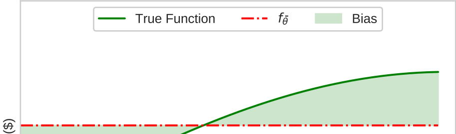
$$\text{Bias}(x) = f_{\text{true}}(x) - E[\hat{f}(x)]$$

**Systematic error:** Difference between truth and expected prediction

**Definition: Variance**

$$\text{Variance}(x) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

**Prediction instability:** Expected squared deviation from mean prediction



# The Main Theorem: Bias-Variance Decomposition

**Important: The Fundamental Result**

**For any point $x$ and any learning algorithm:**

$$E[(y - \hat{f}(x))^2] = \sigma^2 + [\text{Bias}(x)]^2 + \text{Variance}(x)$$

**Definition: Component Interpretation**

- $\sigma^2$: **Irreducible error** (noise)
- $[\text{Bias}(x)]^2$: **Systematic error** (underfitting)
- $\text{Variance}(x)$: **Random error** (overfitting)

**Key Points**

**Coming Up:** We'll prove this step-by-step using careful algebraic manipulation

# Starting the Proof: Expected Squared Error

**Definition: What We're Proving**

$$E[(y - \hat{f}(x))^2] = \sigma^2 + [\mathrm{Bias}(x)]^2 + \mathrm{Variance}(x)$$

**Key Points**

Our Strategy

1. Start with squared error: $(y - \hat{f}(x))^2$
2. Add and subtract strategic terms
3. Expand and use linearity of expectation
4. Show cross-terms cancel out
5. Identify the three components

**Example: Key Insight**

# Step 1: Setting Up the Expectation

## Definition: Squared Loss at Point $x$

**Individual prediction error:** $(y - \hat{f}(x))^2$

## Key Points

Taking Expectations **Expected error over all randomness:**

$$E_{\mathcal{D},y}[(y - \hat{f}(x))^2]$$

where:

- $\mathcal{D}$: Random training set
- $y$: Random target (includes noise)

## Example: Why Two Sources of Randomness?

# Step 2: The Add-and-Subtract Trick

**Key Points**

Starting Point

$$E[(y - \hat{f}(x))^2]$$

**Example: Strategic Addition and Subtraction**

**Add and subtract** $f_{\text{true}}(x)$:

$$E[(y - f_{\text{true}}(x) + f_{\text{true}}(x) - \hat{f}(x))^2]$$

**Definition: Grouping Terms**

$$E[\underbrace{(y - f_{\text{true}}(x))}_{\text{noise: } \epsilon} + \underbrace{(f_{\text{true}}(x) - \hat{f}(x))}_{\text{prediction error}}]^2$$

**Important:**

# Step 3: Expanding the Square

**Key Points**

Using $(a+b)^2 = a^2 + 2ab + b^2$

$$E[\epsilon^2 + 2\epsilon(f_{\text{true}}(x) - \hat{f}(x)) + (f_{\text{true}}(x) - \hat{f}(x))^2]$$

**Example: Applying Linearity of Expectation**

$$E[\epsilon^2] + 2E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))] + E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$

**Definition: Let's Call These**

- **Term 1:** $E[\epsilon^2]$ (the noise term)
- **Term 2:** $2E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))]$ (cross-term)
- **Term 3:** $E[(f_{\text{true}}(x) - \hat{f}(x))^2]$ (prediction error)

# Analyzing Term 1: The Noise Component

**Definition: Term 1 Analysis**

$$E[\epsilon^2] = E[(y - f_{\text{true}}(x))^2]$$

**Key Points**

Key Insight **Independence:** The noise $\epsilon$ doesn't depend on our training set!

- Noise is a property of the data generation process
- Training set selection doesn't affect noise level

**Example: Simplification**

$$E[\epsilon^2] = E[(y - f_{\text{true}}(x))^2] = \text{Var}(\epsilon) = \sigma^2$$

# Analyzing Term 2: The Cross-Term

## Definition: Term 2 Analysis

$$2E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))]$$

## Key Points

Key Independence Property **Crucial insight:** $\epsilon$ (noise) is independent of $\hat{f}(x)$ (our prediction)

- Noise occurs in nature, regardless of our model
- Our model $\hat{f}$ depends only on training data
- Training data and future noise are independent

## Example: Using Independence

$$E[\epsilon(f_{\text{true}}(x) - \hat{f}(x))] = E[\epsilon] \cdot E[f_{\text{true}}(x) - \hat{f}(x)]$$

# Analyzing Term 3: The Prediction Error

**Definition: Term 3 Analysis**

$$E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$

**Key Points**

Another Independence **Key insight:** $(f_{\text{true}}(x) - \hat{f}(x))$ doesn't depend on the noise $\epsilon$

- $f_{\text{true}}(x)$ is deterministic
- $\hat{f}(x)$ depends only on training inputs/outputs (not future noise)

**Example: Simplification**

$$E[(f_{\text{true}}(x) - \hat{f}(x))^2] = \text{MSE of prediction}$$

# Interim Summary: Progress So Far

**Definition: What We Have**

$$E[(y - \hat{f}(x))^2] = \sigma^2 + 0 + E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$
$$= \sigma^2 + E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$

**Key Points**

Next Challenge **Goal:** Decompose $E[(f_{\text{true}}(x) - \hat{f}(x))^2]$ into bias$^2$ + variance

**Example: Strategy for Next Step**

**Another add-and-subtract trick:** We'll add and subtract $E[\hat{f}(x)]$ inside the MSE term

# Step 4: Decomposing the MSE Term

**Key Points**

Current Status

$$E[(y - \hat{f}(x))^2] = \sigma^2 + E[(f_{\text{true}}(x) - \hat{f}(x))^2]$$

**Definition: The Second Add-and-Subtract Trick**

**Add and subtract** $E[\hat{f}(x)]$ in the MSE term:

$$E[(f_{\text{true}}(x) - E[\hat{f}(x)] + E[\hat{f}(x)] - \hat{f}(x))^2]$$

**Example: Grouping Strategically**

$$E[\underbrace{(f_{\text{true}}(x) - E[\hat{f}(x)])}_{\text{bias}} + \underbrace{(E[\hat{f}(x)] - \hat{f}(x))}_{\text{variance}}]^2$$

# Step 5: Expanding the Bias-Variance Square

**Key Points**

Our Expression

$$E[(\text{bias} + \text{variance deviation})^2]$$

**Definition: Let's Define**

- $\alpha = f_{\text{true}}(x) - E[\hat{f}(x)]$ (bias)
- $\beta = E[\hat{f}(x)] - \hat{f}(x)$ (variance deviation)

**Example: Expanding $(\alpha + \beta)^2$**

$$E[(\alpha + \beta)^2] = E[\alpha^2] + 2E[\alpha\beta] + E[\beta^2]$$

# Analyzing the Bias Term: $E[\alpha^2]$

**Definition: The $\alpha^2$ Term**

$$E[\alpha^2] = E[(f_{\text{true}}(x) - E[\hat{f}(x)])^2]$$

**Key Points**

Key Insight $(f_{\textbf{true}}(x) - E[\hat{f}(x)])$ **is deterministic!**

- $f_{\text{true}}(x)$ is a fixed function value
- $E[\hat{f}(x)]$ is the expected prediction (a constant)
- No randomness left to average over!

**Example: Simplification**

$$E[\alpha^2] = (f_{\text{true}}(x) - E[\hat{f}(x)])^2 = [\text{Bias}(x)]^2$$

# Analyzing the Cross-Term: $E[\alpha\beta]$

**Definition: The Cross-Term**

$$E[\alpha\beta] = E[(f_{\text{true}}(x) - E[\hat{f}(x)])(E[\hat{f}(x)] - \hat{f}(x))]$$

**Key Points**

Another Independence Argument **Key insight:** $(f_{\text{true}}(x) - E[\hat{f}(x)])$ is deterministic (constant)

- Can factor it out of the expectation
- Remaining term has expectation zero

**Example: Factoring Out the Constant**

$$E[\alpha\beta] = (f_{\text{true}}(x) - E[\hat{f}(x)]) \cdot E[E[\hat{f}(x)] - \hat{f}(x)]$$
$$= \text{bias} \cdot (E[\hat{f}(x)] - E[\hat{f}(x)]) = \text{bias} \cdot 0 = 0$$

# Analyzing the Variance Term: $E[\beta^2]$

**Definition: The $\beta^2$ Term**

$$E[\beta^2] = E[(E[\hat{f}(x)] - \hat{f}(x))^2]$$

**Key Points**

Recognizing the Pattern **This is exactly the definition of variance!**

$$\text{Variance}(\hat{f}(x)) = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

**Example: Rewriting for Clarity**

$$E[\beta^2] = E[(E[\hat{f}(x)] - \hat{f}(x))^2] = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$
$$= \text{Variance}(\hat{f}(x))$$

**Important: Result**

# The Complete Bias-Variance Decomposition

**Important: Putting It All Together**

$$E[(y - \hat{f}(x))^2] = \sigma^2 + [\text{Bias}(x)]^2 + \text{Variance}(x)$$

**Definition: Component Summary**

- $\sigma^2$ = **Irreducible error** (noise in data)
- $[\text{Bias}(x)]^2$ = **Systematic error** (model assumptions)
- $\text{Variance}(x)$ = **Random error** (training set sensitivity)

**Key Points**

The Fundamental Tradeoff

- **Reduce bias:** Use more complex models → Increase variance
- **Reduce variance:** Use simpler models → Increase bias

# Summary: The Bias-Variance Tradeoff

**Definition: What We've Proven**

**Every prediction error can be decomposed as:**

$$\text{Total Error} = \text{Noise} + \text{Bias}^2 + \text{Variance}$$

**Key Points**

Key Takeaways

- **Noise:** Cannot be reduced (irreducible)
- **Bias:** Reduced by increasing model complexity
- **Variance:** Reduced by decreasing model complexity
- **Optimal model:** Balances bias and variance

**Important: Practical Applications**